Ficha de Unidade Curricular

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular *(1.000 carateres).*

Técnicas estatísticas para análise de mineração de dados / Statistical techniques for data mining analysis

1.2. Sigla da área científica em que se insere *(100 carateres).*

MAT

1.3. Duração[1] *(100 carateres).*

Semestral

1.4. Horas de trabalho[2] *(100 carateres).*

162

1.5. Horas de contacto[3] *(100 carateres).*

TP: 45; PL:22.5

1.6. ECTS. *(100 carateres).*

6

1.7. Observações[4]. *(1.000 carateres).*

1.7. Remarks. *(1.000 carateres).*

2. Docente responsável e respetiva carga letiva na Unidade Curricular. *(preencher o nome completo) (1.000 carateres).*

Sandra Maria da Silva Figueiredo Aleixo (22.5h)

3. Outros docentes e respetivas cargas letivas na unidade curricular *(1.000 carateres).*

Iola Maria Silvério Pinto (22.5h)
Carlos José Brás Geraldes (22.5h)

4. Objetivos de aprendizagem (conhecimentos, aptidões e competências a desenvolver pelos estudantes) *(1.000 carateres)*.

A UC introduz algoritmos e métodos estatísticos para mineração de dados. Alia estatística, bases de dados e computação. Os objetivos são:

1. Rever conceitos de Probabilidade e Estatística. Identificar as fases de um projeto de Ciência de Dados (PCD). Aprender conceitos e técnicas estatísticas de aprendizagem automática (AA)
2. Identificar o tipo de dados e aprender métodos para a sua preparação e pré-processamento
3. Planear e implementar uma base de dados segundo o modelo relacional
4. Aplicar técnicas para transformação no âmbito dos modelos de regressão nomeadamente modelos lineares generalizados, modelos aditivos generalizados e redes neuronais
5. Conhecer a base matemática dos métodos de AA apresentados, saber manuseá-los, identificar e interpretar várias formas e tipos de resultados

6. Avaliar os resultados obtidos com as técnicas de AA, usando e interpretando as medidas de desempenho
7. Usar softwares apropriados
8. Realizar um PCD no âmbito da regressão, com metodologias adequadas

4. Intended learning outcomes (knowledge, skills and competences to be developed by the students) ***(1.000 characters)*****.**

The CU introduces statistical data mining algorithms and methodologies. It combines statistics, databases and computation. The intended outcomes are:

1. Review concepts of Probability and Statistics. To describe the several stages of a Data Science project (DSP). To know the concepts and statistical techniques of machine learning (ML)
2. Identify the type of data and to know the methodologies for its preparation and preprocessing
3. Planning and implement a data base accordingly to the relational model
4. Apply techniques for transformation in the regression models context, such as generalized linear model, generalized additive model and neural networks
5. Know the mathematical basis of the ML methods presented, know how to handle them, identify and interpret various forms and types of results
6. Evaluate the results obtained with the ML techniques, using and interpreting the performance measures
7. Use appropriate software
8. Complete a DSP under the regression, with appropriate methodologies

5. Conteúdos programáticos ***(1.000 carateres).***

1. Revisão de Probabilidade e Estatística.
2. Introdução à Mineração de Dados. Projeto de Mineração de Dados.
3. Dados: Tipos; Qualidade; Pré-processamento.
4. Bases de Dados: Modelo Entidade Relacionamento; Implementação do Modelo Físico; Consultas.
5. Métodos Lineares para Regressão - Ridge e Lasso.
6. Modelos Lineares Generalizados. Modelos de Regressão Logística.
7. Análise de Sobrevivência.
8. Métodos Não Lineares para Regressão.
9. Modelos Aditivos Generalizados.
10. Modelos de Regressão baseados em Redes Neuronais.

5. Syllabus ***(1.000 characters).***

1. Review of Probability and Statistics.
2. Introduction to Data Mining. Data Mining Project.
3. Data: Type; Quality; Pré-processing.
4. Data Bases: The Entity-Relationship Model; Physical Model implementation; Queries.
5. Regression Linear Methods - Ridge e Lasso.
6. Generalized Linear Models. Logístic Regression Models.
7. Survival Analysis.
8. Non Linear Regression Methods.

9. Generalized Additive Models.
10. Regression models based on Neural Networks.

6. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular ***(1.000 carateres).***

Os pontos 1 e 2 dos conteúdos programáticos pretendem alcançar o ponto 1 dos objetivos
O ponto 3 dos conteúdos programáticos introduz os conceitos necessários para atingir o ponto 2 dos objetivos
O ponto 4 dos conteúdos programáticos pretende alcançar o ponto 3 dos objetivos
Os restantes pontos (5 a 10) dos conteúdos programáticos pretendem alcançar os restantes objetivos (4 a 8)

6. Evidence of the syllabus coherence with the curricular unit's intended learning outcomes ***(1.000 characters).***

Topics 1 and 2 of the syllabus aims to achieve the topic 1 of the objectives
Topic 3 of the syllabus introduces the concepts necessary to achieve the objectives topic 2
Topic 4 of the syllabus aims to achieve the topic 3 of the objectives
The rest of syllabus topics (5 until 10) focuses on the consolidation of goals 4 until 8

7. Metodologias de ensino (avaliação incluída)*)* ***(1.000 carateres).***

As aulas são teórico-práticas. É utilizada uma metodologia expositiva para a apresentação da matéria teórica, exemplificada com a resolução de exercícios e de problemas concretos, implementada computacionalmente usando a linguagem R.
A avaliação de conhecimentos compreende duas componentes, uma teórica (NT) e outra prática (NP). A componente teórica é constituída por um exame (nota mínima de 9,5 valores). A componente prática é constituída por um trabalho de grupo (nota mínima de 9,5 valores). Este trabalho será um dos casos de estudo apresentado de entre vários, deve ser desenvolvido ao longo do semestre e deve integrar as várias fases de um projeto de Ciência de Dados.
A nota final do aluno (NF) será obtida através da fórmula NF=0,4NT+0,6NP.

7. Teaching methodologies (including assessment) ***(1.000 characters).***

The classes are theoretical-practical. An expository methodology is used for the presentation of the theoretical matter, exemplified with the resolution of exercises and concrete problems, implemented in a computational way using R language.
The knowledge assessment comprises two components, one theoretical (TG) and another practical (PG). The theoretical component consists of an exam (minimum grade of 9.5 points). The practical component consists of a group work (minimum grade of 9.5 values). This work will be one of the case studies presented among several, should be developed throughout the semester and should integrate the various phases of a Data Science project.
The student's final grade (FG) will be obtained through the formula FG = 0.4TG + 0.6PG.

8. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular ***(3.000 carateres).***

As metodologias de ensino são coerentes com os objetivos da unidade curricular, dado que a metodologia expositiva utilizada para explicar a matéria teórica possibilita atingir os objetivos da unidade curricular. A utilização de exemplos resolvidos computacionalmente associados aos diversos tópicos do programa, permite dotar os alunos de competências adequadas para

a resolução dos desafios colocados pelo tecido empresarial no mercado de trabalho. A elaboração de um projeto completo de Ciência de Dados, usando metodologias apropriadas, será uma mais-valia para quando os alunos iniciarem a sua atividade profissional.
O método de avaliação permite averiguar se o aluno adquiriu os conhecimentos necessários para atingir os objetivos propostos na unidade curricular.

8. Evidence of the teaching methodologies coherence with the curricular unit's intended learning outcomes. ***(3.000 characters).***

The teaching methodologies are consistent with the objectives of the curricular unit, given that the expository methodology used to explain the theoretical subject makes it possible to reach the objectives of the curricular unit. The use of computationally solved examples, associated with the various topics of the program allows students to have adequate skills to solve the challenges posed by the job market. The implementation of a complete Data Science project, using appropriate methodologies, will be an advantage when students begin their professional activity.
The evaluation method allows to verify if the student has acquired the necessary knowledge to reach the objectives proposed in the curricular unit.

9. Bibliografia de consulta/existência obrigatória ***(1.000 carateres).***

James, G., Witten, D., Hastie, T. and Tibshirani, R., An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics (2017).

1. Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. Springer (2017).
2. Bishop, C. M., Pattern Recognition and Machine Learning, Springer (2006)
3. Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V., Introduction to Data Mining, Pearson (2019)
4. Witten, I. H., Frank, E. and Hall M. A., Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2011).
5. Hand, D. J., Mannila, H. and Smyth, P., Principles of data mining. The MIT Press (2001).
6. Torgo, L., Data mining with R – learning with case studies. CRC Press (2010).
7. Zhao, Y., R and Data Minning: Examples and Case Studies, Elsevier (2012).
8. Lantz, B., Machine Learning with R, Packt (2013).
9. Muller, A. and Guido, S., Introduction to Machine Learning with Python, O'Reilly (2017).
10. Murphy, K., Machine Learning: A Probabilistic Perspective, MIT Press (2012).

---

[1] Anual, semestral, trimestral, …
[2] Número total de horas de trabalho.
[3] Discriminadas por tipo de metodologia adotado (T - Ensino teórico; TP - Ensino teórico-prático; PL - Ensino prático e laboratorial; TC - Trabalho de campo; S - Seminário; E - Estágio; OT - Orientação tutorial; O - Outro).
[4] Assinalar sempre que a unidade curricular seja optativa.